Accelerated oblique random survival forests

Byron C. Jaeger

Last updated 2022-06-14

Overview

- Random forests (axis based and oblique)
  - Decision trees
  - Random survival forests (RSF)
- Accelerating the oblique RSF
  - Newton Raphson scoring
- Benchmark
  - Datasets & learners
  - Evaluation
  - Results
- Software
  - R package & website

Random forests

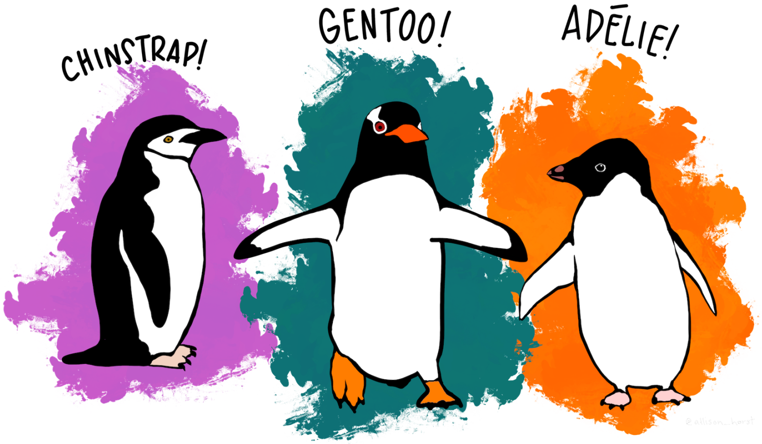
(axis based and oblique)

Decision trees

- Frequently used in supervised learning.
- Partitions the space of predictor variables.
- Can be used for classification, regression, and survival analysis.

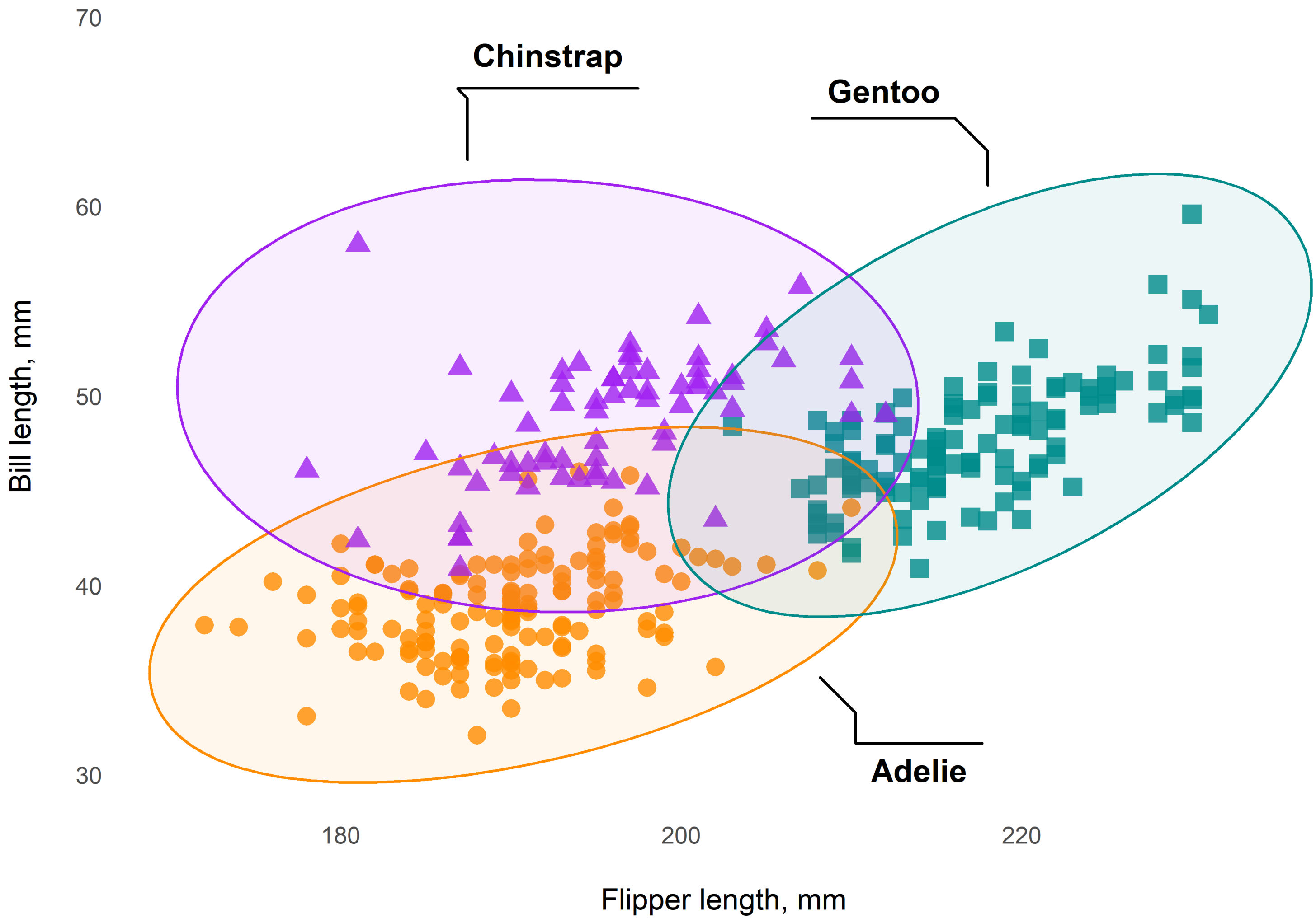
Demo:

Axis-based and oblique decision trees for classification of penguin species (chinstrap, gentoo, or adelie) based on bill and flipper length.[1]

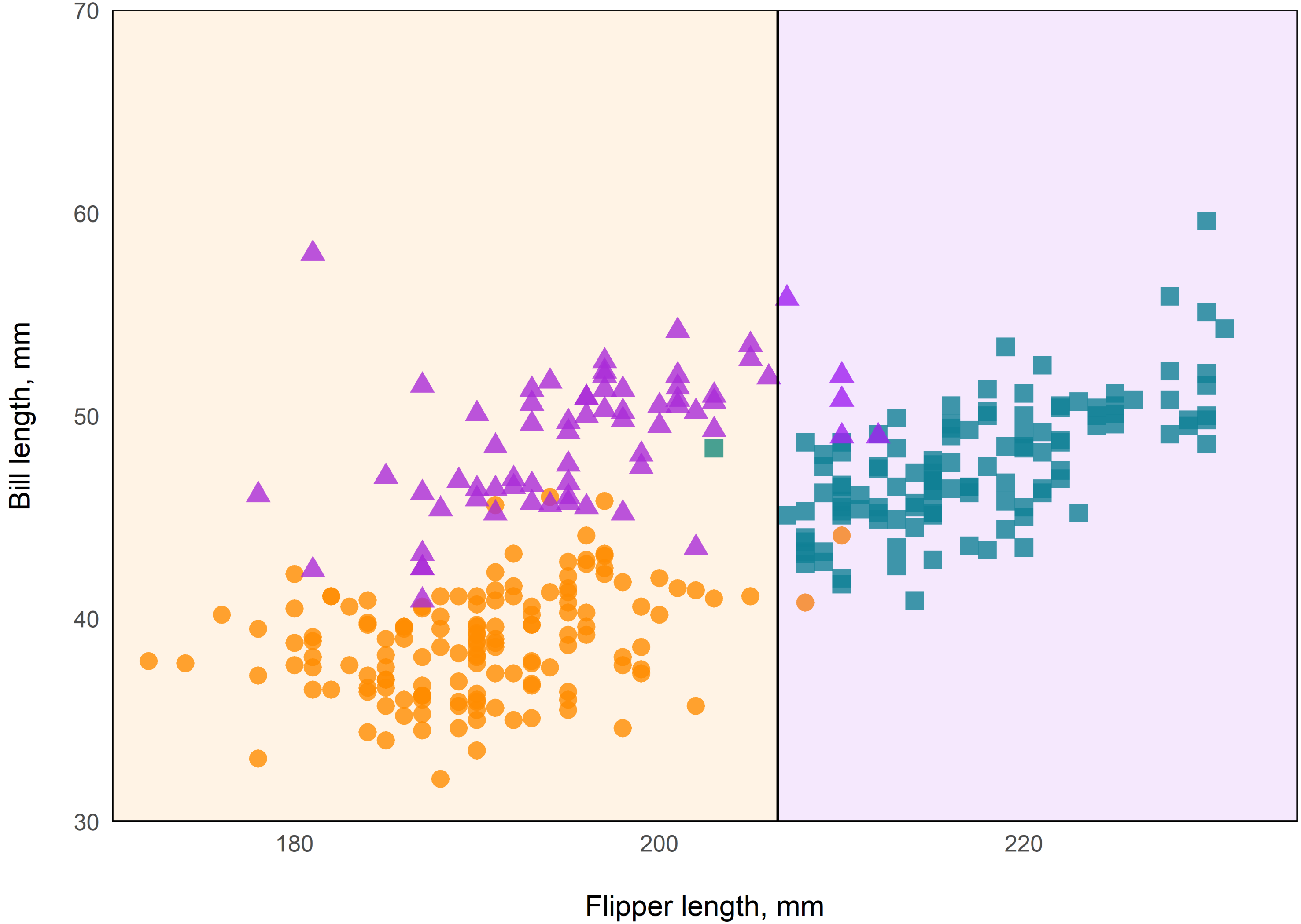


[1]Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, a member of the Long Term Ecological Research Network.

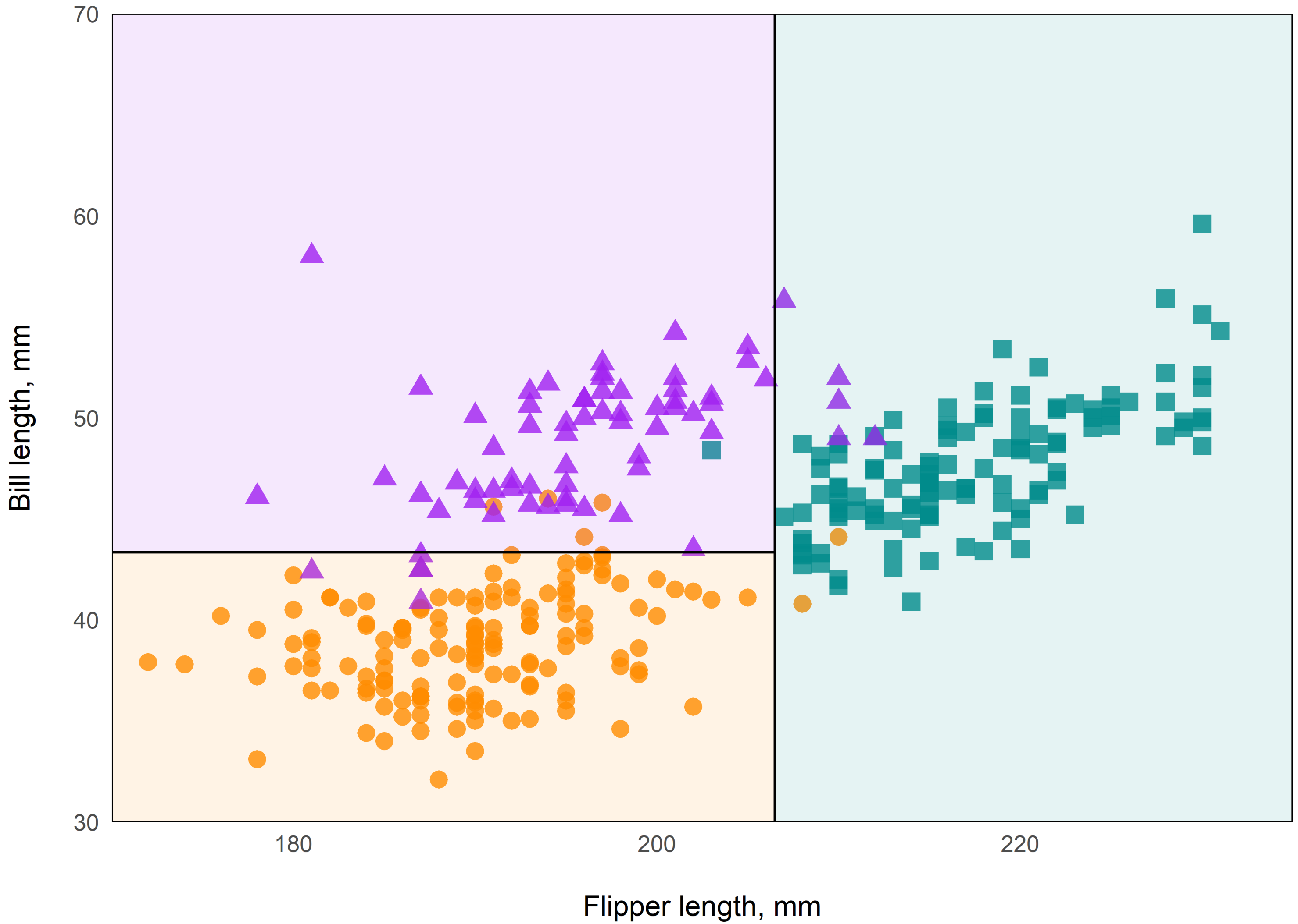
Dimensions for Adelie, Chinstrap and Gentoo Penguins at Palmer Station



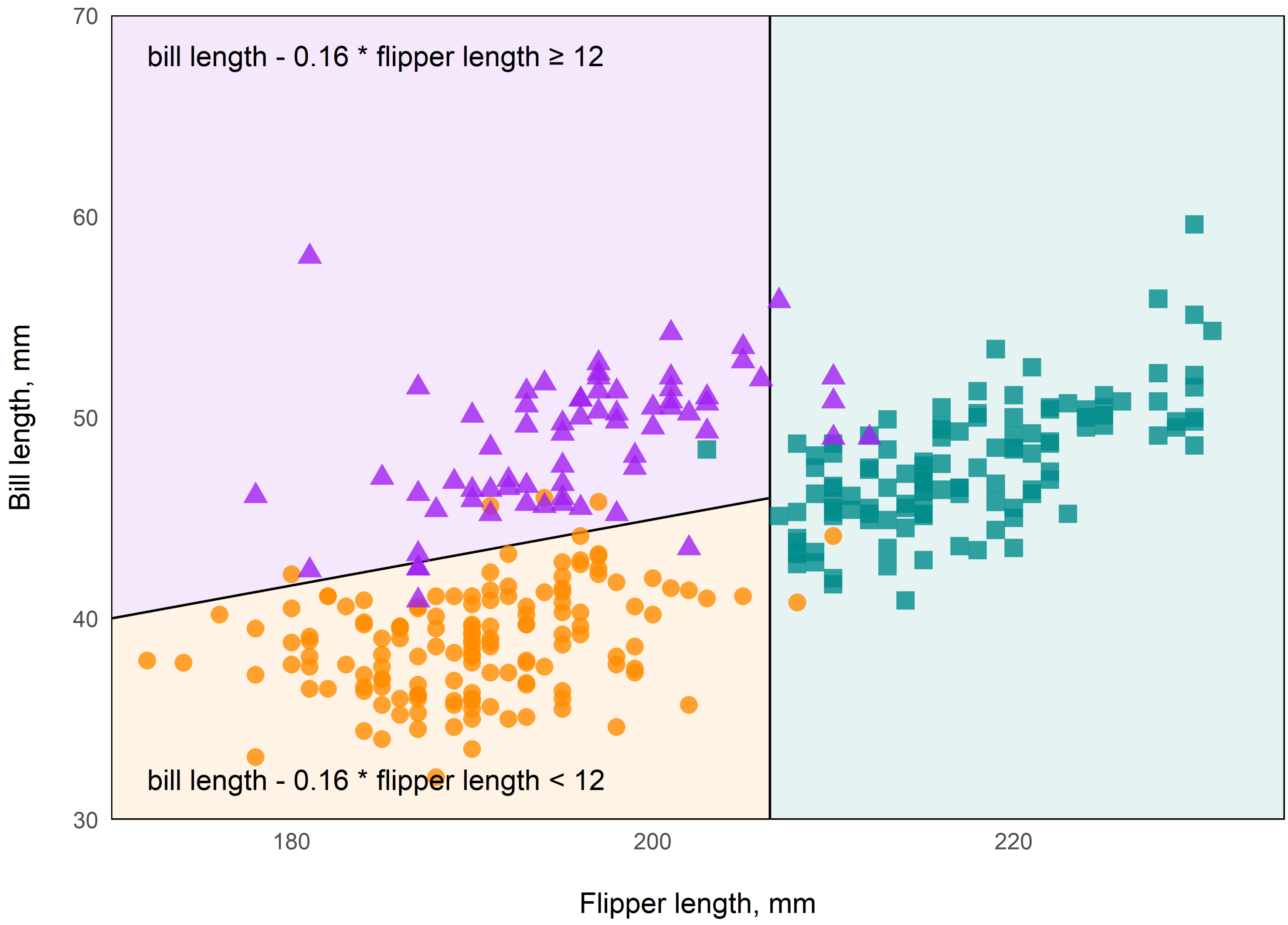
Partition all the penguins into flipper length < 207 or ≥ 207 mm



Partition penguins on the left side into into bill length < 43 or ≥ 43 mm



With oblique splits, partitions do not need to be rectangles



Random survival forests (RSFs)

1. Breiman developed the random forest, a large set of decision trees injected with randomness.[1, 2]

[1] Breiman, Leo. "Bagging predictors." Machine learning 24.2 (1996): 123-140.
[2] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

Random survival forests (RSFs)

1. Breiman developed the random forest, a large set of decision trees injected with randomness.[1, 2]

2. Hothorn and, separately, Ishwaran developed extensions of the random forest for survival outcomes.[3, 4]

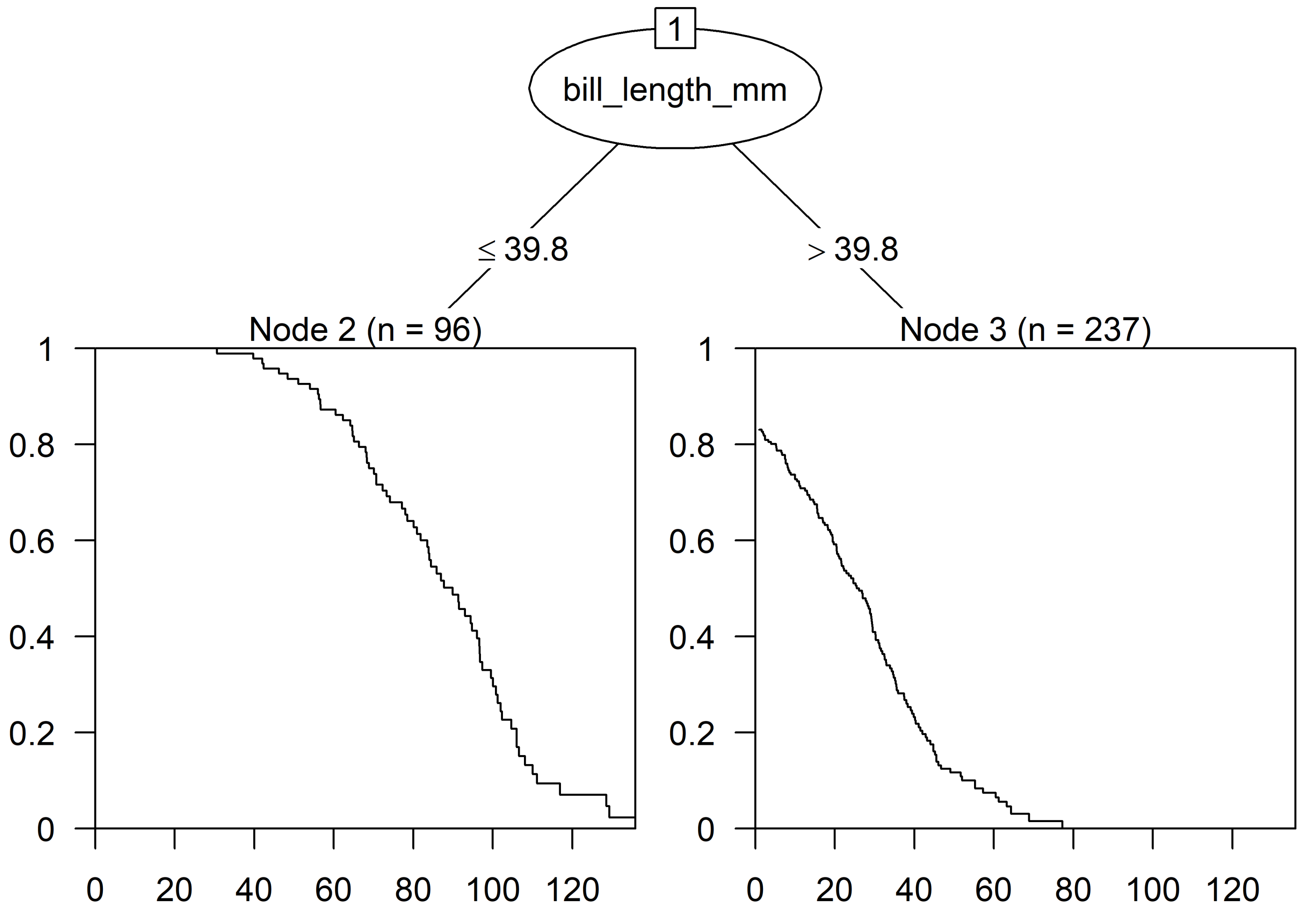
Specifically,

- Hothorn developed the conditional inference forest (CIF)

- Ishwaran developed the random survival forest (RSF)

[3]Hothorn, Torsten, et al. "Unbiased recursive partitioning: A conditional inference framework." Journal of Computational and Graphical statistics 15.3 (2006): 651-674.
[4] Ishwaran, Hemant, et al. "Random survival forests." Annals of Applied Statistics 2.3 (2008): 841-860.

Each leaf in the RSF contains a Kaplan-Meier estimate of survival. In the CIF, weights are applied based on sample size.



Random survival forests (RSFs)

1. Breiman developed the random forest, a large set of decision trees injected with randomness.[1, 2]

2. Hothorn and, separately, Ishwaran developed extensions of the random forest to engage with survival outcomes (CIF and RSF).[3, 4]

3. Zhou developed a rotation survival forest and Wang developed a survival forest with an extended predictor space.[5, 6]

Both Zhou and Wang's extensions were based on the CIF

[5]Zhou L, et al. "Rotation survival forest for right censored data." PeerJ. 2015 Jun 11;3:e1009.
[6] Wang H, et al. "Random survival forest with space extensions for censored data." Artif Intell Med. 2017 Jun;79:52-61.

Random survival forests (RSFs)

1. Breiman developed the random forest, a large set of decision trees injected with randomness.[1, 2]

2. Hothorn and, separately, Ishwaran developed extensions of the random forest to engage with survival outcomes (CIF and RSF).[3, 4]

3. Zhou developed a rotation survival forest and Wang developed a survival forest with an extended predictor based on the CIF.[5, 6]

4. Jaeger developed the oblique RSF, which used penalized regression to find oblique splits.[7]

Jaeger showed in general benchmarks that the oblique RSF had higher prediction accuracy than axis-based RSFs. However, the obliqueRSF R package is hundreds of times slower than standard R packages for axis based RSFs.

[7]Jaeger BC, et al. "Oblique random survival forests." The Annals of Applied Statistics 13.3 (2019): 1847-1883.


Tell me the truth...I'm...I'm ready
to hear it.
NOBODY USES SLOW R PACKAGES
imgflip.com

Accelerating the oblique RSF

Accelerating the oblique RSF

We identify linear combinations of predictor variables in non-leaf nodes by applying Newton Raphson scoring to the partial likelihood function of the Cox regression model:

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{x_{j(i)}^{'}\beta}}{\sum_{j \in R_i} e^{x_j^{'}\beta}}$$

- x_i is a vector of predictors values.
- R_i is the set of indices, j, with $T_j \geq t_i$ (i.e., those still at risk at time t_i)
    - T_i is the event time if an event occurred and last point of contact otherwise.
    - $t_1 < \ldots < t_m$ are the m unique event times in the training data.
- $j(i)$ is the index of the observation for which an event occurred at time t_i.

Newton Raphson scoring

Estimated regression coefficients $\hat{\beta}$ are updated in each step based on their first derivative, $U(\hat{\beta})$, and second derivative, $H(\hat{\beta})$:

$$\hat{\beta}^{k+1} = \hat{\beta}^{k} + U(\hat{\beta} = \hat{\beta}^{k})\, H^{-1}(\hat{\beta} = \hat{\beta}^{k})$$

For statistical inference, iterate until a convergence threshold is met.

For identifying linear combination of predictors in the oblique RSF, 🤷‍♀️

- `aorsf-fast` completes one iteration.
- `aorsf-cph` iterates until convergence or 15 iterations.

Benchmark

Learners

Oblique RSFs:

- *aorsf-fast*: the fast version of aorsf
- *aorsf-cph*: the less fast but still pretty fast version of aorsf
- *aorsf-random*: randomized coefficients (Breiman's idea)
- *aorsf-net*: aorsf's copy of obliqueRSF
- *obliqueRSF-net*: oblique RSF using penalized cox regression (the original)

Learners

Oblique RSFs:

- *aorsf-fast*: this is the only oblique RSF we show in the racing model plots.
- *aorsf-cph*: the less fast but still pretty fast version of aorsf
- *aorsf-random*: randomized coefficients (Breiman's idea)
- *aorsf-net*: aorsf's copy of obliqueRSF
- *obliqueRSF-net*: oblique RSF using penalized cox regression (the original)

Learners

Axis based RSFs:

- *cif-standard*: standard CIF
- *cif-extension*: CIF with space extension
- *cif-rotate*: CIF with rotation
- *rsf-standard*: standard RSF
- *ranger-extratrees*: RSF with extremely randomized trees

Learners

Other:

- *glmnet-cox*: Penalized Cox regression model
- *nn-cox*: Cox neural network with time-varying effects
- *xgboost-cox*: Boosted trees fitted to Cox log likelihood
- *xgboost-aft*: Boosted trees fitted to accelerated failure time.

Data sets

A total of 23 risk prediction tasks in 16 data sets were analyzed.

- number of observations ranged from 137 to 17549 (median = 1151)
- number of predictors ranged from 7 to 1692 (median = 12)
- % censored ranged from 5 to 98 (median = 68)

(Full table is shown in the bonus slides)

Evaluation

We measured performance of each learner with:

- index of prediction accuracy (IPA); **higher** is 👍
- time-dependent concordance (C)-statistic; **higher** is 👍
- total time to fit a model and compute predictions; **lower** is 👍

Evaluation

To estimate overall performance differences:

Step 1: Collect IPA and C-statistic values:

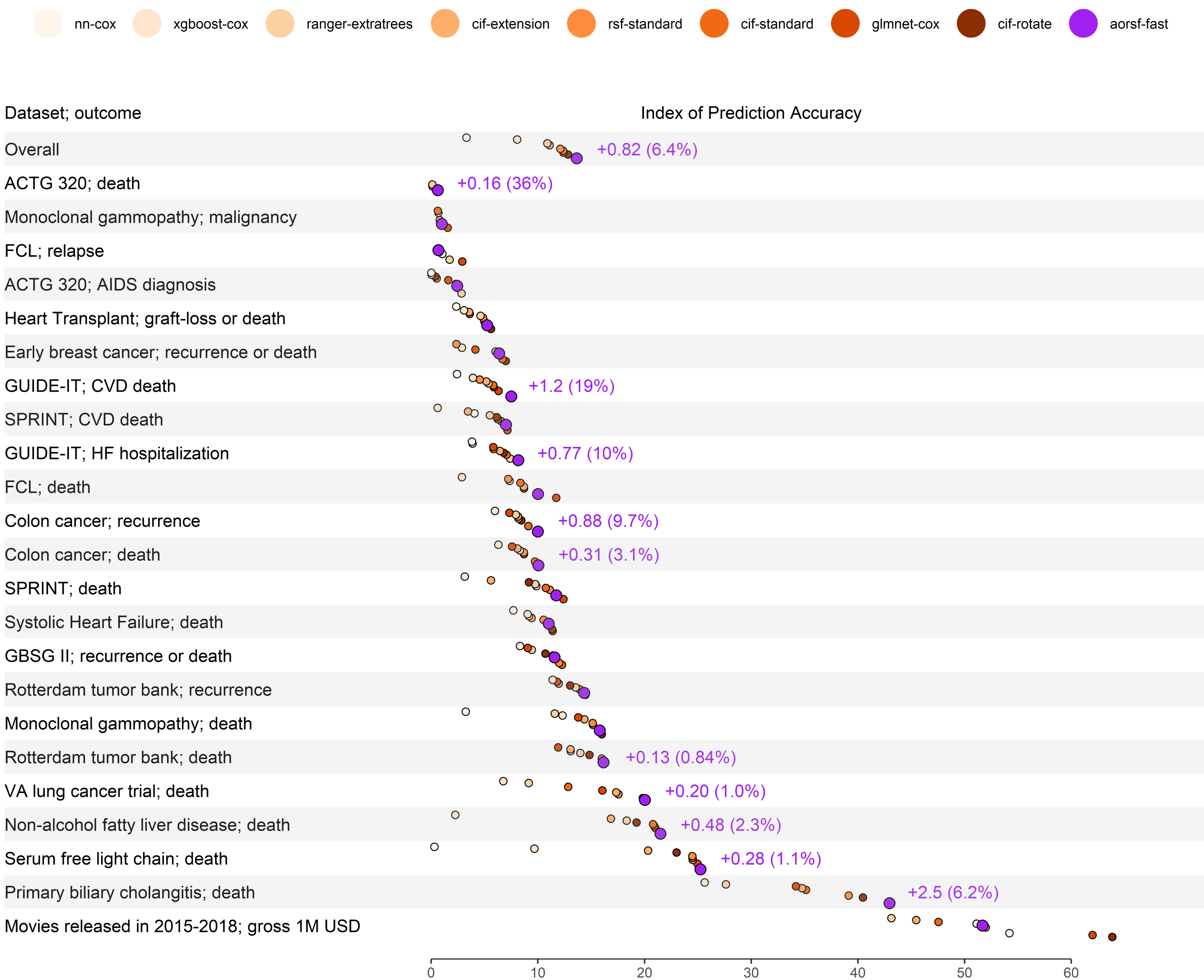
- For each of the 23 risk prediction problems,
    - split the corresponding data into a 50/50 train/test split
    - fit each learner to the training set
    - evaluate each learner's predictions in the testing set
    - repeat 25 times

Step 2: Fit a hierarchical Bayesian model to analyze posterior expected differences in performance:

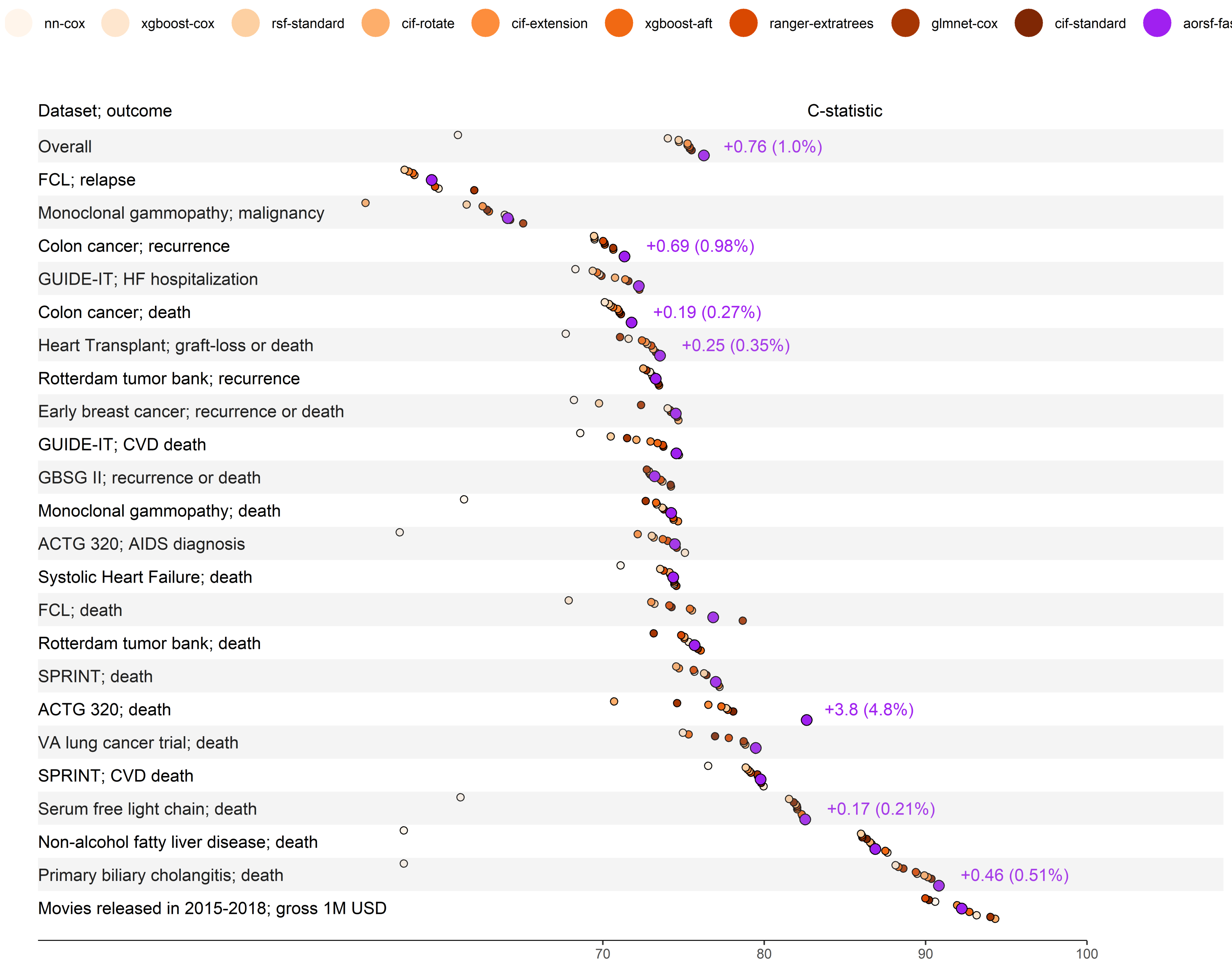
$$\text{metric} = \widehat{\gamma} \cdot \text{model} + (1 \,|\, \text{data}/\text{run})$$

- `run` refers to the specific train/test split of `data`

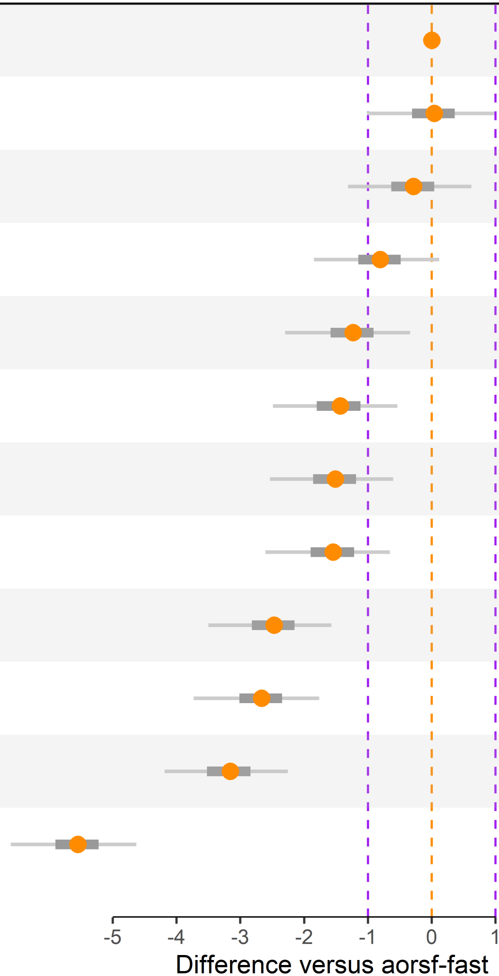
- `metric` is either the IPA or the time-dependent C-statistic.



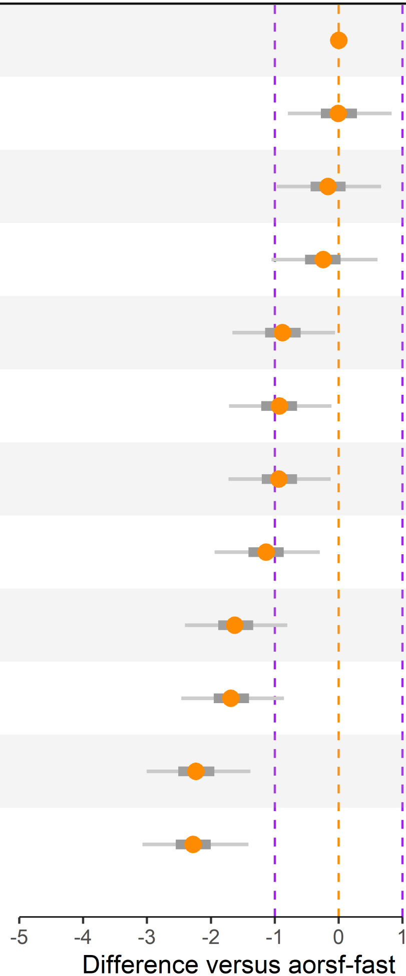
nn-cox
xgboost-cox
ranger-extratrees
cif-extension
rsf-standard
cif-standard
glmnet-cox
cif-rotate
aorsf-fast
Dataset; outcome
Index of Prediction Accuracy
Overall
+0.82 (6.4%)
ACTG 320; death
+0.16 (36%)
Monoclonal gammopathy; malignancy
FCL; relapse
ACTG 320; AIDS diagnosis
Heart Transplant; graft-loss or death
Early breast cancer; recurrence or death
GUIDE-IT; CVD death
+1.2 (19%)
SPRINT; CVD death
GUIDE-IT; HF hospitalization
+0.77 (10%)
FCL; death
Colon cancer; recurrence
+0.88 (9.7%)
Colon cancer; death
+0.31 (3.1%)
SPRINT; death
Systolic Heart Failure; death
GBSG II; recurrence or death
Rotterdam tumor bank; recurrence
Monoclonal gammopathy; death
Rotterdam tumor bank; death
+0.13 (0.84%)
VA lung cancer trial; death
+0.20 (1.0%)
Non-alcohol fatty liver disease; death
+0.48 (2.3%)
Serum free light chain; death
+0.28 (1.1%)
Primary biliary cholangitis; death
+2.5 (6.2%)
Movies released in 2015-2018; gross 1M USD
0
10
20
30
40
50
60

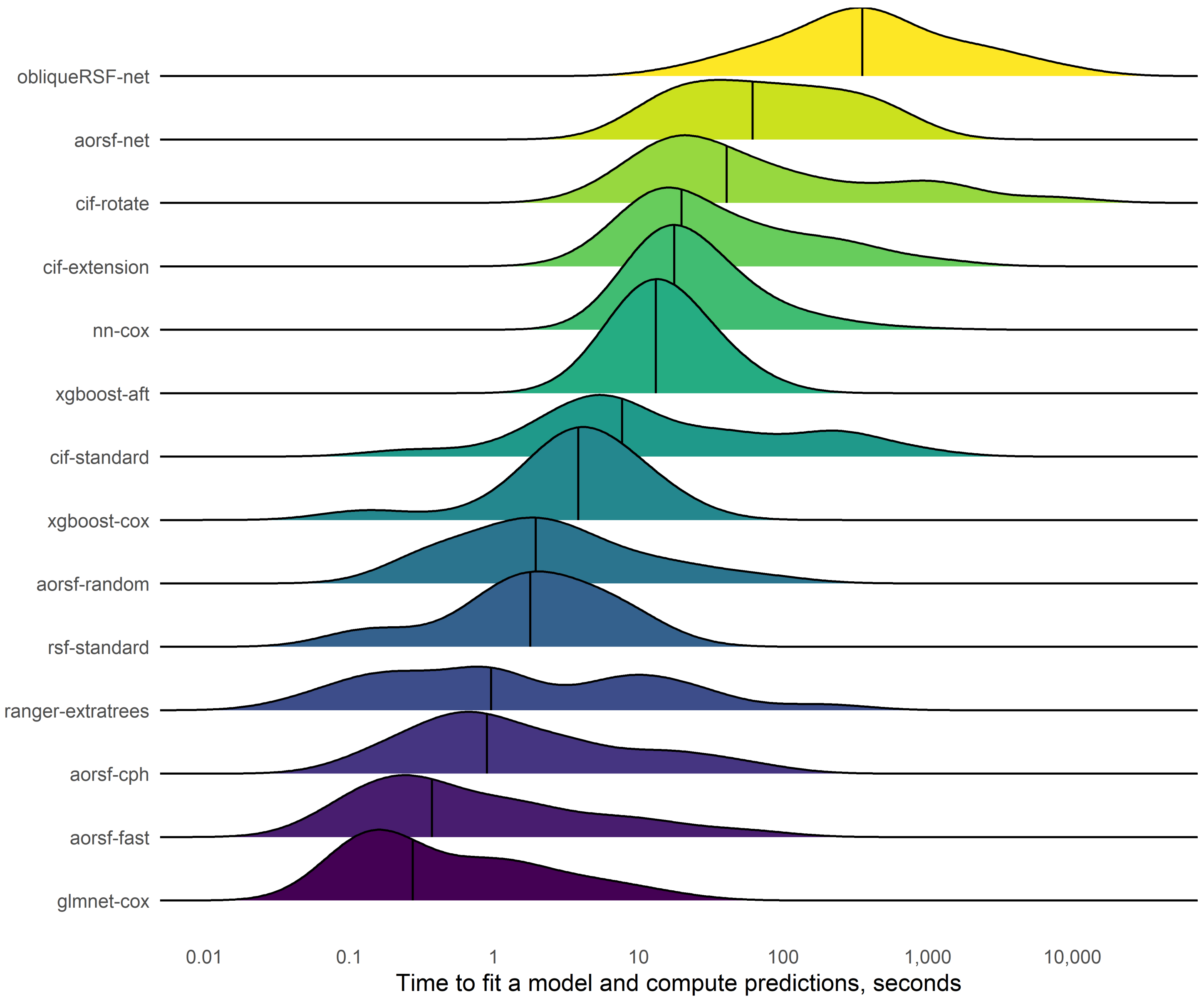

nn-cox
xgboost-cox
rsf-standard
cif-rotate
cif-extension
xgboost-aft
ranger-extratrees
glmnet-cox
cif-standard
aorsf-fa
Dataset; outcome
C-statistic
Overall
+0.76 (1.0%)
FCL; relapse
Monoclonal gammopathy; malignancy
Colon cancer; recurrence
+0.69 (0.98%)
GUIDE-IT; HF hospitalization
Colon cancer; death
+0.19 (0.27%)
Heart Transplant; graft-loss or death
+0.25 (0.35%)
Rotterdam tumor bank; recurrence
Early breast cancer; recurrence or death
GUIDE-IT; CVD death
GBSG II; recurrence or death
Monoclonal gammopathy; death
ACTG 320; AIDS diagnosis
Systolic Heart Failure; death
FCL; death
Rotterdam tumor bank; death
SPRINT; death
ACTG 320; death
+3.8 (4.8%)
VA lung cancer trial; death
SPRINT; CVD death
Serum free light chain; death
+0.17 (0.21%)
Non-alcohol fatty liver disease; death
Primary biliary cholangitis; death
+0.46 (0.51%)
Movies released in 2015-2018; gross 1M USD
70
80
90
100

		Posterior probability		
Learner	Scaled integrated Brier score	Equivalence	Difference < 0	Difference < -1
aorsf-fast		---	---	---
aorsf-cph		0.95	0.46	0.03
aorsf-net		0.91	0.72	0.09
cif-rotate		0.65	0.96	0.35
cif-standard		0.31	1.00	0.69
glmnet-cox		0.17	1.00	0.83
rsf-standard		0.15	1.00	0.85
obliqueRSF-net		0.13	1.00	0.87
cif-extension		0.00	1.00	1.00
ranger-extratrees		0.00	1.00	1.00
aorsf-random		0.00	1.00	1.00
xgboost-cox		0.00	1.00	1.00



Learner	Time-dependent C-statistic	Posterior probability		
		Equivalence	Difference < 0	Difference < -1
aorsf-fast		---	---	---
aorsf-cph		0.98	0.51	0.00
aorsf-net		0.98	0.66	0.02
obliqueRSF-net		0.96	0.72	0.03
cif-standard		0.62	0.98	0.38
ranger-extratrees		0.56	0.99	0.44
glmnet-cox		0.57	0.99	0.43
cif-extension		0.37	1.00	0.63
rsf-standard		0.07	1.00	0.93
cif-rotate		0.05	1.00	0.95
xgboost-cox		0.00	1.00	1.00
aorsf-random		0.00	1.00	1.00




obliqueRSF-net
aorsf-net
cif-rotate
cif-extension
nn-cox
xgboost-aft
cif-standard
xgboost-cox
aorsf-random
rsf-standard
ranger-extratrees
aorsf-cph
aorsf-fast
glmnet-cox
0.01
0.1
1
10
100
1,000
10,000
Time to fit a model and compute predictions, seconds

Using the accelerated oblique RSF

aorsf 0.0.0.9000

aorsf

`aorsf` provides optimized software to fit, interpret, and make predictions with oblique random survival forests (ORSFs).

Why aorsf?

- over 400 times faster than `obliqueRSF`.
- accurate predictions for time-to-event outcomes.

Thank you!

BONUS ROUND

Data sets

A total of 23 risk prediction tasks in 16 data sets were analyzed.

This table is continued on the next 2 slides.

	Outcome	N Obs	N Predictors	N Events	% Censored	% Missing	% Continuous
ACTG 320	AIDS diagnosis	1,151	12	96	92	0.00	30
ACTG 320	Death	1,151	12	26	98	0.00	30
Colon cancer	Recurrence	929	12	468	50	0.37	20
Colon cancer	Death	929	12	452	51	0.37	20
Early breast cancer	Recurrence or death	614	1,692	134	78	0.00	100
FCL	Death	541	7	76	86	0.00	40
FCL	Relapse	541	7	272	50	0.00	40

Data sets continued

	Outcome	N Obs	N Predictors	N Events	% Censored	% Missing	% Continuous
GBSG II	Recurrence or death	686	10	299	56	0.00	62
GUIDE-IT	CVD death	894	59	110	88	12	47
GUIDE-IT	HF hospitalization	894	59	288	68	12	47
Heart Transplant	Graft-loss or death	3,787	52	500	87	6.1	26
Monoclonal gammopathy	Death	1,384	8	963	30	0.49	83
Monoclonal gammopathy	Malignancy	1,384	8	115	92	0.49	83
Movies released in 2015-2018	Gross 1M USD	551	46	522	5.3	0.00	4.5
Non-alcohol fatty liver disease	Death	17,549	24	1,364	92	33	45

Data sets continued

	Outcome	N Obs	N Predictors	N Events	% Censored	% Missing	% Continuous
Primary biliary cholangitis	Death	276	19	111	60	0.00	59
Rotterdam tumor bank	Recurrence	2,982	11	1,518	49	0.00	56
Rotterdam tumor bank	Death	2,982	11	1,272	57	0.00	56
Serum free light chain	Death	7,874	10	2,169	72	1.7	50
SPRINT	CVD death	9,361	174	521	94	0.65	24
SPRINT	Death	9,361	174	1,644	82	0.65	24
Systolic Heart Failure	Death	2,231	41	726	67	0.00	33
VA lung cancer trial	Death	137	8	128	6.6	0.00	33

Evaluation (bonus round)

Consider a testing data set:

$$\mathcal{D}_{\text{test}} = \{(T_i, \delta_i, x_i)\}_{i=1}^{N_{\text{test}}}.$$

Let $\widehat{S}(t \mid x_i)$ be the predicted probability of survival up to a given prediction horizon of $t > 0$. The Brier score at time t is

$$\widehat{\text{BS}}(t) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \Big\{ \widehat{S}(t \mid x_i)^2 \cdot I(T_i \leq t, \delta_i = 1) \cdot \widehat{G}(T_i)^{-1} + \\ [1 - \widehat{S}(t \mid x_i)]^2 \cdot I(T_i > t) \cdot \widehat{G}(t)^{-1} \Big\}$$

where $\widehat{G}(\cdot)$ is the Kaplan-Meier estimate of the censoring distribution.

Evaluation (bonus round)

As the Brier score is time dependent, integration over time provides a summary measure of performance over a range of plausible prediction horizons. The integrated Brier score is defined as

$$\widehat{\mathcal{BS}}(t_a, t_b) = \frac{1}{t_b - t_a} \int_{t_a}^{t_b} \widehat{\mathrm{BS}}(t) dt.$$

In our results

- t_a is the 25th percentile of event times
- t_b is the 75th percentile of event times

Evaluation (bonus round)

$\widehat{\mathcal{BS}}(t_a, t_b)$, a sum of squared prediction errors, can be scaled to produce a measure of explained residual variation (i.e., an R^2 statistic) by computing

$$R^2 = 1 - \frac{\widehat{\mathcal{BS}}(t_a, t_b)}{\widehat{\mathcal{BS}}_0(t_a, t_b)}$$

where $\widehat{\mathcal{BS}}_0(t_a, t_b)$ is the integrated Brier score when a Kaplan-Meier estimate for survival based on the training data is used as the survival prediction function $\widehat{S}(t)$.

Jargon: R^2 = **index of prediction accuracy** (IPA)